A Machine Learning Approach to Predicting SEP Proton Intensity and Events Using Time Series of Relativistic Electron Measurements

Jesse Torres[1], Philip K. Chan[1], Lulu Zhao[3], Ming Zhang[2]

[1]Department of Computer Engineering and Sciences, Florida Institute of Technology, Melbourne, FL 32901, USA
[2]Department of Aerospace, Physics and Space Sciences, Florida Institute of Technology, Melbourne, FL 32901, USA
[3]Department of Climate and Space Sciences and Engineering, University of Michigan, Ann Arbor, MI 48108, USA

Key Points:

- We use neural network models to predict proton flux 30 and 60 minutes into the future from measurements of energetic electrons.
- Neural network models can improve the prediction accuracy of proton intensity and SEP events.

Corresponding author: Jesse Torres, jtorres2015@my.fit.edu

Abstract

Solar energetic particles (SEPs) can cause severe damage to astronauts and sensitive equipment in space, and can disrupt communications on Earth. A lack of thorough understanding the eruption processes of solar activities and the subsequent acceleration and transport processes of energetic particles makes it difficult for physics-based models to forecast the occurrence of an SEP event and its intensity. Therefore, in order to provide an advance warning for astronauts to seek shelter in a timely manner, we apply neural networks to forecast the intensity of SEP events. The neural network uses a time series of past and current electron and proton flux in 5-minute intervals to predict future proton flux 30 minutes or 1 hour ahead. In addition to regular neural networks, we also use recurrent neural networks (RNNs), which are designed to handle time series data. For each model, we consider two approaches: a single model trained on all data, and the ensemble of models where the particular model is selected dynamically for each input using the predicted behavior of the input data. Overall, our results indicate that a single RNN model forecasts proton flux of each event with less error. Furthermore, the RNN model incurs less error in predicting proton flux, but a larger lag, than the forecasting matrix method proposed by Posner (2007). When advance and extended warnings are incorporated, the RNN model can improve SEP event prediction scores.

Plain Language Summary

One possible threat to the astronauts and the equipment during any space exploration mission is the high-energy radiation coming from Solar Energetic Particle (SEP) events. SEP events cannot be prevented, so astronauts must be given advance warning so they have enough time to seek shelter. The goal of this research is to forecast the high-energy proton radiation environment in the near-Earth space in a continuous manner. We use neural networks, a method which can learn patterns from a dataset and then use the learned model to make predictions. The neural network outputs proton flux (SEP intensity) ahead in time, for example by 30 minutes or 1 hour, with a 5-minute cadence of every SEP electron and proton flux measurement. The machine learning uses a stream of past and current electron and proton fluxes as well as other information of solar eruptions to predict future fluxes of high-energy SEP protons. Relativistic electrons travel faster and arrive at Earth earlier than sub-relativistic protons, so they could serve as an advance warning of the arrival of more lethal sub-relativistic protons. The machine learning model results show an improved prediction of SEP proton intensity and event occurrence.

1 Introduction

Solar energetic particles (SEPs) are high-energy particles from the Sun which, at a high enough intensity, can cause harm to astronauts and sensitive equipment in space. The intensity of these events is measured in proton flux. A solar eruption can accelerate SEPs up to tens of GeV in larger events, and the flux of >10 MeV protons could increase many orders of magnitude ($>10^4$) over the background level from Galactic cosmic rays. Protons of >150 MeV are very difficult to shield, and they can penetrate 20 gm cm^{-2} of material, i.e., 7.4 cm of Al or 15.5 cm of water (or human flesh) (Reames, 2013). Because of the danger presented by intense SEP events, it is essential to provide an advance warning so that astronauts can move themselves and their equipment to safety timely.

SEPs contain many species: protons, alpha particles, heavier nucleons, electrons, and X-rays. High-energy protons are the biggest concern of all. They are most abundant, difficult to shield, and can deposit all their energies along their tracks deep into human tissue or electronics on spacecraft. Relativistic electrons typically arrive at Earth earlier than sub-relativistic protons because electrons travel faster and experience less scat-

tering during their propagation through the interplanetary medium. Observations show that near relativistic electrons in the energy range from a few hundred keV to a few MeV appear in almost every event when high-energy SEP protons are observed (Posner, 2007), and the rise of electron intensity typically precede that of protons by tens of minutes. Therefore, measurements of relativistic electrons can become an advanced signal for the arrival of more lethal protons. This work focuses on the prediction of SEP proton intensity using leading information from near relativistic SEP electrons.

The predicted intensity is proton flux, which can also be measured continuously over time. Therefore, we use time series data for features which precede the event in order to predict a time series of proton flux. Specifically, a fixed window of time containing past and current values of proton and electron flux in measurement cadence intervals (5 minutes in this study) are used for input, and a value of proton flux is predicted at either half an hour or one hour in the future. We compare the performance of the standard neural network implementation (NN) with recurrent neural networks (RNN), which contain additional weight matrices characterizing dependencies between inputs across time, making them a good fit for time series forecasting problems. An additional contribution of our approach is the separation of data into different intensity ranges, which are used to train and test multiple models. This approach is meant to address the imbalance between SEP events and background values.

The rest of this paper is organized as follows: Section 2 discusses previous works related to our studies. Section 3 describes our approach, including the input and output and the algorithm. Section 4 goes into detail on the experimental evaluation, including the data, evaluation criteria, procedures, results, and analysis. Finally, Section 5 summarizes the paper and discusses limitations and potential improvements.

2 Related Work

Related studies have 3 general directions in forecasting. The first direction uses properties of solar flares. Most of the existing prediction models use either the post-eruptive observations of solar flares to forecast or nowcast SEPs (NOAA-SWPC-PROTON (Balch, 2008); AFRL-PPS (Smart & Shea, 1976, 1989; Huang et al., 2012; Belov, 2009; Laurenza et al., 2009) or the forecast of solar flares from the sun's magnetic field measurement (Georgoulis, 2008; E. Park et al., 2018; Bobra & Ilonidis, 2016; Huang et al., 2018) to forecast SEPs. Among similar works, Garcia (2004) performs a spectral analysis of hard x-ray bursts, and uses them to determine whether an SEP will occur based on the hardening of the spectral index. Then, they predict the magnitude of the event by applying an empirical function using the max temperature and max x-ray flux. In the work from Kahler et al. (2005), the Proton Prediction System (PPS) uses peak x-ray flux and the x-ray flare rise time in order to forecast peak proton flux. If the predicted proton flux exceeds the 10 pfu threshold, which indicates that there is an SEP event, then the location of the solar flare is used by the PPS to predict the onset and peak times of the SEP. Núñez (2011) uses two models: one for well-connected events and one for poorly-connected events. In events when there is a direct magnetic connection between Earth and solar flare, the first time derivatives of x-ray and proton fluxes are found to be correlated. The model for the well-connected events can use the correlation and the associated solar flare to predict SEP proton event. For poorly-connected events, an ensemble of model trees is used for predicting future proton flux of a poorly-connected event using past proton flux.

The second direction of SEP prediction studies utilizes CME properties. Compared to using the properties of solar flares to predict SEPs, it is less popular to use CMEs to predict SEPs due to the fact that the identification of CMEs requires image processing and recognition from observations in application to real-time SEP prediction. However, a potential nowcast of SEPs is reported by St. Cyr et al. (2017) using the near-real-time temporal cadence (15 s) of K-cor observations very close to the solar disk ($1.05-3R_s$)

in the Mauna Loa Solar Observatory (MLSO). During the 2016 January 1 SEP event, MLSO was the first to issue warning comparing to all other predicting techniques (St. Cyr et al., 2017). An empirical formula is developed by Richardson et al. (2018) which uses the speed and connection angle of CMEs to predict the peak intensity of 14- to 24-MeV protons. The formula tends to overpredict the intensity for small, predominantly western events, since the formula was obtained based on large wide-spread three-spacecraft events. Although this formula predicts intensity, it can be used to forecast SEP occurrence by using a threshold of proton flux 10^{-4} (MeV s cm^2 sr)$^{-1}$. The prediction algorithm is evaluated using the false alarm rate (FAR) and probability of detection (POD), among other metrics, and with different portions of the data depending on which combinations of type II solar radio emissions are present. Although the CME speed shows positive correlation with the occurrence and intensity of SEPs (J. Park et al., 2012; Dierckxsens et al., 2015; Cane et al., 2010), for a given CME speed, the SEP intensity can vary over 3 orders of magnitude. Moreover, a slow CME could drive a stronger shock with a mach number of 3.43−4.18 and a fast CME could drive a weaker shock with a mach number of 1.90 − 3.21 (Shen et al., 2007). It is not clear how SEP intensity relates to the CME speed, given that there are many other factors that could affect SEP events.

The third direction is to use measurements of relativistic electrons as an advanced signal. Posner (2007) found relativistic electron intensity onset to occur earlier than proton intensity by up to one hour ahead. They use time series data of 1-minute increments. Each instance consists of the current measurements of electron intensity and the max rate of electron rise (difference in consecutive electron measurements) within a window from 5 minutes ago to 1 hour ago. The model consists of a 18x13 matrix, in which the rows are ranges of 0.3-1.2 MeV electron intensity values and the columns are ranges of electron rise rate values. The predicted proton intensity in each cell is the average future (up to 1 hour ahead) 30-50 MeV proton intensity of all instances within that cell's input parameter ranges. Forecasting is performed by finding the corresponding matrix cell for a given test instance, and predicting the proton intensity in that cell.

3 Approach

We follow the basic concept of the predictive power contained in relativistic electron measurements. We propose using a time series of past and current electron and proton intensity measurements and their time derivatives to forecast proton intensity ahead of time. Instead of just using the definitive correlation between proton intensity and electron intensity and its time derivative in Posner (2007), we use a machine learning approach to the problem. This way, we can build a model that can include subtle correlations between future proton intensity and all other available time series measurements. We also make different models according to conditions based on the radiation level.

In Section 3.1, we discuss the time series and features in more detail. In Section 3.2, the basic machine learning approach is discussed, and extensions to this approach using multiple models are detailed in Section 3.3.

3.1 Input and Output for Forecasting Proton Intensity

We use time series of particle measurements to predict future >10 MeV proton intensity. The learning and prediction are made based on the natural log scale of particle intensity in pfu. A time window of the past two hours up to the current hour is used to predict an output proton flux at 30 minutes or 1 hour from the current time. The cadence of input data is 5 minutes. Let t be the current time, t+1 to be 5 minutes after the current time, and t-1 to be 5 minutes before the current time. Then the input time window is from t-24 through t, and the output is either t+6 or t+12.

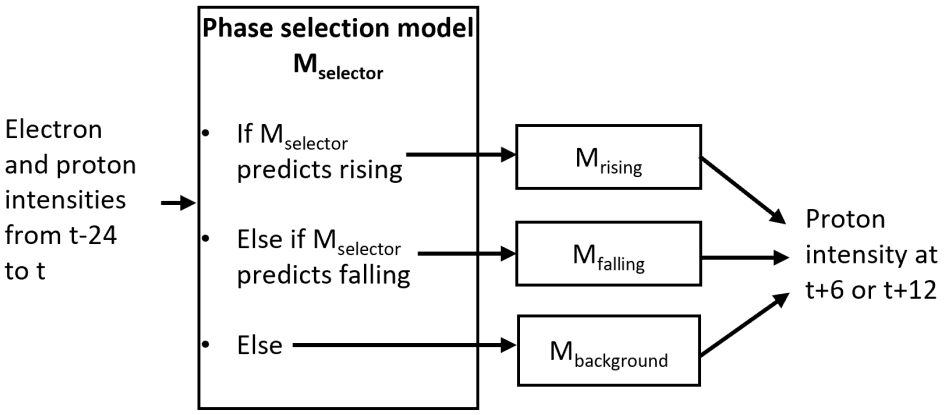


Figure 1. Illustration of the method used for selecting one of multiple models to use for proton intensity prediction.

The main features used are the electron intensities from the >0.25 and >0.67 MeV channels and the proton intensities from the >10 MeV channel obtained by the EPHIN instrument on *SOHO* at the L1 point (Müller-Mellin et al., 1995). The >10 MeV proton intensity is calculated from the intensities in three differential energy (P8, P25, and P41) channels through integration of power-law extrapolated/interpolated spectrum. The data set covers 1995-2002.

In addition to the above features, we add phases to the feature input for each timestep of the past 2 hours in order to help the model to handle different stages of SEP events. We define three phases for an SEP event: rising, falling, and background. These phases are determined by a separate program designed for identifying the onset, threshold, peak, and end timestamps of each event. Using these timestamps, we can determine when the intensity is rising (between onset and peak), falling (between peak and end), and background (everywhere else), and use these as the phase inputs. However, when an event occurs in practice, it is not known that the intensity is rising or falling until some time after the transition. Therefore, for any change of phase in the input, we use the previous phase as the label until 30 minutes pass, after which we replace them with the true phase. This is done for both the training and testing sets.

3.2 Basic Approach with One Model

To form a baseline, we apply a single neural network model to the whole dataset; this approach will be referred to as M1 (One Model for intensity prediction). We compare the multilayer perceptron algorithm (which from here on will be denoted as NN, or neural network) to recurrent neural networks (RNNs); recurrent neural networks are designed to work with time-series data, so they are expected to perform better in this study. The *Keras* implementation in Python is used to create our models, with the Gated Recurrent Units (*GRU*) layer for the recurrent neural network model. GRU layers are described in more detail by Chung et al. (2014). During training, the neural network minimizes the loss function, which is the mean square error (MSE) between the target proton intensity and the predicted proton intensity.

3.3 Multiple Models

Due to the substantial imbalance between background flux and SEP events, the basic approach of training a single neural network on all of the data could miss most of

the rare SEP events. Methods such as oversampling of the SEP events could address this issue, but are computationally expensive when the training set is large. Instead of training a single model, we train multiple models. These models are specialized for certain situations so that they can be more accurate. The approach involves two training stages; the first training stage is model selection, and the second training stage is proton flux prediction using multiple models. We refer to this approach as M3 (Three Models for intensity prediction).

Figure 1 illustrates the process for predicting the proton intensity via selecting an intensity model based on the predicted phase. D is the training set with features and known proton intensity. D_{phase} is the training set containing all the same features as D, but with a known phase. We use the program mentioned in Section 3.1 in order to create background, rising, and falling labels for D_{phase}. The first stage of training is to train the model selector $M_{selector}$, a classifier model that is trained on D_{phase} to predict the phase of a training instance. $M_{selector}$ makes predictions on D (in which the phase is not known), and based on these predictions, D is split into disjoint subsets $D_{background}$, D_{rising}, and $D_{falling}$. In the second stage of training, the three proton intensity models $M_{background}$, M_{rising}, and $M_{falling}$ are trained on their respective training sets which were determined in the first training stage. Testing is performed by using $M_{selector}$ to determine whether a test instance is to be passed to $M_{background}$, M_{rising}, or $M_{falling}$, then using the chosen model to predict proton intensity.

Since model selection (phase prediction in our case) is a classification task, categorical cross-entropy is used as the loss function for the model selector. Since the data is imbalanced and we cannot split it for this model, we set class weights such that there is a one-to-one weight ratio between falling and background instances, and a three-to-one ratio between rising and background. Additionally, we experiment with setting the sample weights to be 4 times higher than other instances for all samples between the onset of an event and when the intensity reaches ln(10). Larger weights are given to the rising class in order to help the algorithm with on-time prediction of the onset through the rising edge. Further detail on weighting procedures is described in Section 4.2.1.

4 Experimental Evaluation

We evaluate our proposed methods with existing methods on two forecasting tasks. The first task is forecasting the proton intensity 30 minutes (t+6) and 60 minutes (t+12) ahead. The second task is forecasting the occurrence of SEP events (proton intensity exceeding a threshold) 30 minutes and 60 minutes ahead.

4.1 Evaluation Criteria

4.1.1 Evaluation of SEP Time-Intensity Profile Prediction

The intensity models are primarily evaluated using mean absolute error (MAE). This calculates the absolute difference between the actual and predicted values at each timestamp. Since we are only interested in the SEP events, we calculate MAE only for the events found by the event-finding program that occur within the test set, and average over all of these events. For the purposes of evaluation, we consider an event to be the rising portion, from onset to peak. In Figure 2, the vertical teal arrow shows the error at one timestamp; this is calculated for all timestamps between the onset and the peak and averaged to obtain MAE.

Another factor we want to look at is whether the predictions are on-time or too late. To do this, we measure lag by shifting the predictions between the onset and peak of each event in the test set to align with the targets, and measuring MAE at each shift. The shift with the lowest MAE is considered the lag, or x-axis error, between the tar-

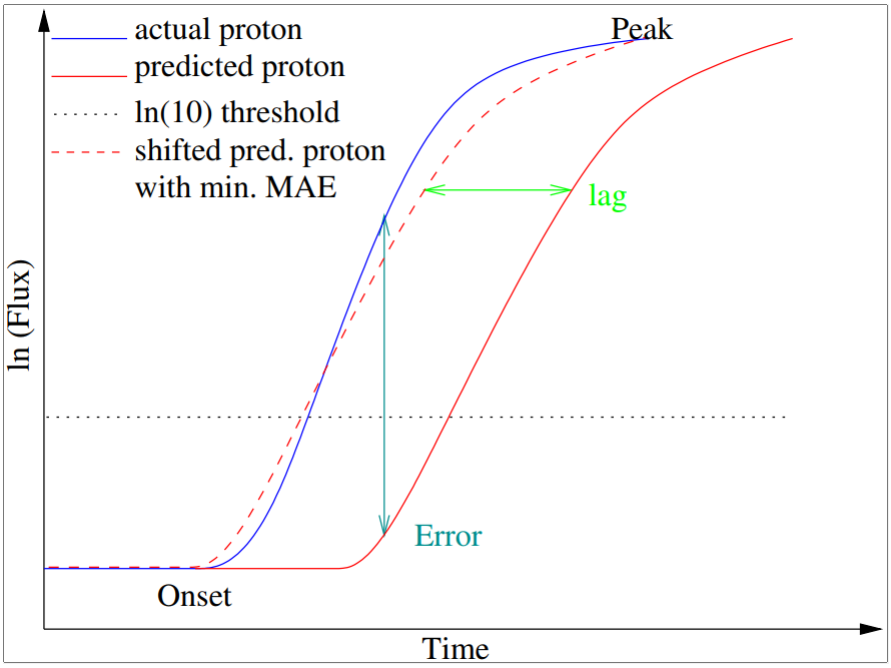


Figure 2. A diagram to illustrate the metrics for evaluating intensity predictions

gets and predictions. In Figure 2, the procedure can be visualized as incrementally shifting the solid red line left and measuring MAE between the red and blue lines at each shift, with the dashed red line being the lag with the minimum MAE. Similarly with MAE, this is done only for the events in the test set rather than the entire test set, and between the onset and the peak.

4.1.2 Evaluation of SEP Event Forecasting

In addition to evaluating the proton intensity predictions in terms of error and lag, we also assess the approach's ability to forecast SEP events; that is, predicting the occurrence of >10 MeV proton intensity above the 10 pfu threshold before the actual intensity crosses the threshold. In order to perform this evaluation, some terms must be defined first. We define the warning period as all consecutive timestamps during which the prediction is above the threshold. The warning period of an alert is represented by the red bars in Figure 3. Alert 1 of Figure 3 is an example where the warning is able to detect SEP event 1, as the warning is ongoing at the time that SEP event 1 starts. However, the warning alone is not always sufficient for event detection. Alert 2 is an example of Issue 1, where a gap in the warning period can lead to the event being missed; these gaps occur when the predictions fluctuate below the threshold. Alert 3 is an example of Issue 2, in which the event can also be missed if the warning ends before the start of the event. In order to address these two issues, we prolong the warning after the prediction drops below the threshold; the prolonged warning is called the *extended warning* period, indicated by the green bars in Figure 3. The extended warning can be considered as smoothing out the fluctuations in the forecast. To investigate an appropriate duration for the extended warning, we vary the duration from 15 minutes up to 2.5 hours in increments of 15 minutes. The extended warning may end before the specified duration if the prediction exceeds the threshold again.

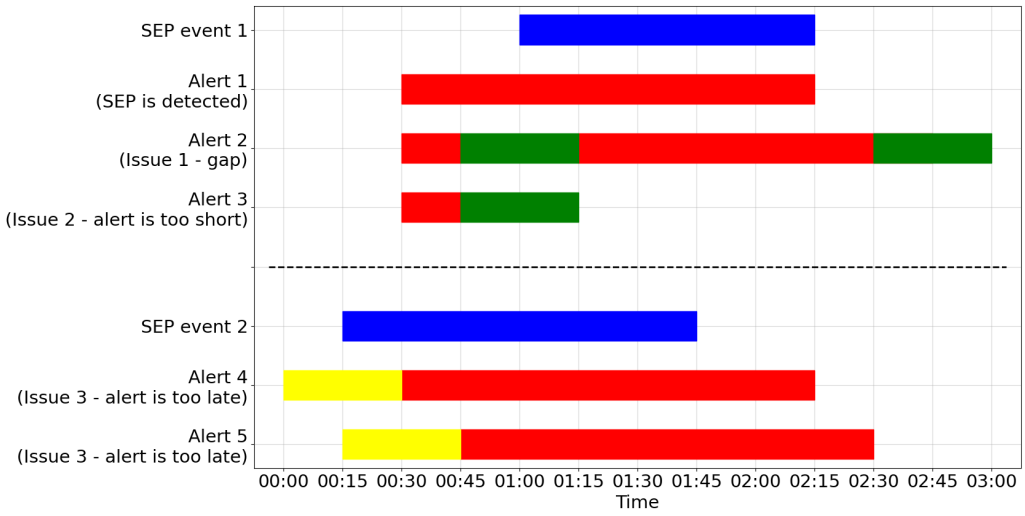


Figure 3. Different scenarios of SEP events (blue) and warnings (red); issues of event detection which can be addressed by extended warning (green) and advance warning (yellow). In this example, the extended warning lasts 30 minutes.

Issue 3 is that the warning can be too late, as shown in Alert 4 of Figure 3, where the warning begins after SEP event 2 starts. Since the proton intensity predictions for t+6 or t+12 are generated at time t, we can address this issue by beginning to warn at time t; the period between time t and either t+6 or t+12 is called the *advance warning* period, indicated by the yellow bars in the figure. Alert 5 is another example of Issue 3, in which the warning is too late and the advance warning is unable to forecast the SEP event. We will further discuss this alert in Section 4.3 to illustrate the weakness of the persistent model, one of the baseline models for comparative evaluation.

To investigate the merits of advance and extended warning, we evaluate four different approaches in which different combinations of advance and extended warning are used to define an alert. An approach which uses neither is used as a baseline for comparison. The warning period, during which the prediction is above the threshold, is always included in the alert. The four approaches are as follows:

- Approach W (baseline): without advance or extended warning
- Approach EW: with extended warning, but without advance warning
- Approach AW: with advance warning, but without extended warning
- Approach EAW: with both advance and extended warnings

The performances of the four approaches are evaluated in terms of true positives, false negatives, and false positives. An alert is a true positive (TP) if it starts before an event starts and ends after the event starts. An event is a false negative (FN) if either there is no alert corresponding to the event, or there is an alert, but it starts after the event has begun (i.e. the alert is too late). Finally, an alert is a false positive (FP) if it has no overlap with any event duration. True negatives cannot be counted since the absence of both alerts and events cannot be quantified. Additionally, composite metrics

including recall, precision, and F1 score (the harmonic mean of recall and precision) are calculated; these are defined in Equations 1, 2, and 3.

$$recall = \frac{TP}{TP + FN} \tag{1}$$

$$precision = \frac{TP}{TP + FP} \tag{2}$$

$$F1 = \frac{2 * precision * recall}{precision + recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{3}$$

4.2 Learning and Evaluation Procedures

The training and testing sets are split chronologically such that the first 80% of the usable data is for training and the last 20% is for testing. (Note on usable data: since each instance requires 2 hours of data before it and either 30 minutes or 1 hour after, the first 24 and the last 6 or 12 timestamps are unusable as the current time.) There are 21 SEP events in the training set, and 18 SEP events in the test set. In each of the three approaches, the features used are the >0.25 MeV electron intensities, >0.67 MeV electron intensities, and >10 MeV proton intensities, and all are tested with and without phase inputs, with both NN and RNN algorithms.

For each of the intensity models, a single hidden layer with 30 units is used; this layer is changed from fully-connected to recurrent for the RNN experiments. For both regular and recurrent neural networks, the hidden layer uses a sigmoid activation. Weight updates are done using the Adam optimizer, and up to 1000 iterations are allowed unless the network converges before then. The neural network converges if the loss function does not change by more than 10^{-4} over 20 iterations. Each experiment is run five times with different random initialization, and the average and standard deviation of MAE and lag are reported for each metric.

4.2.1 Phase-Selection Model Procedures

In order to incorporate the machine learning phase-selection model into the M3 method, some weighting must be applied to address the class imbalance between rising instances and other instances. In a preliminary experiment, the phase-selection model was tested without class weights, with class weights, and with sample weights for the NN algorithm; RNN was only tested with class weights, which was determined to be the best at distinguishing background from rising instances by the time we tested with RNNs. Class weights are assigned to be three times higher for rising instances than for other instances, and sample weights are assigned to be four times higher for instances between event onsets and ln(10) thresholds than for other instances. When testing with the RNN, both the phase-selection model and the intensity models are RNNs.

To evaluate the machine learning-based phase selection model, we look at a 3x3 confusion matrix for the three classes of background, rising, and falling. We emphasize on-time classification of the rising class since the main goal is to identify the start of events, so we look at precision and, more importantly, recall for the rising class. We also look at the F1 score, which is the harmonic mean of the precision and recall. To summarize the results of the model selector described by Torres (2020), it was found that adding weights to rising instances when training the model selector always yielded improvement over not using weights in terms of intensity prediction performance. The addition of sample weights yielded slightly better performance than only using class weights, but using only class weights resulted in fewer errors in instances where the prediction is background

but the actual instance is rising. That is, only using class weights results in fewer SEP events missed, and so the phase-selection model in this paper only uses class weights.

4.2.2 Baseline Methods Used for Comparison

To further assess the performance of our approaches, we compare our models with two baseline methods. The first is the *persistent model*, a simple baseline method in which the predicted proton flux value is the current proton flux value, which is equivalent to nowcast. The second is the forecasting matrix method described by Posner (2007) as described in Section 2. SOHO has two energy channels of 0.3-1.2 MeV electron measurements, while the electron intensities we use are >0.25 MeV. Furthermore, the forecasting matrix predicts 30-50 MeV proton intensities, while our methods predict >10 MeV proton intensities, so the matrix in Posner (2007) cannot be used to perform direct comparison. Therefore, we must train and test a forecasting matrix on the dataset we use, with some adjustments to the implementation. In our implementation, we maintain the same number of matrix cells (13x18) and clip the input parameter values such that all cells have at least one instance, and therefore can predict a proton intensity using any cell. These bounds are 0.01 to 0.2 per minute for the time derivation of log of electron intensity, and -3 to 8 for the log of the current electron intensity in pfu. Since both baseline methods are deterministic, the results are reported for only a single run. We evaluate both methods in terms of MAE and lag in the time series , and we compare the results with our neural network approach. Additionally, we evaluate Posner's method's ability to detect events and compare results with the neural network approach.

4.3 Results of Predicting SEP Time-Intensity Profile

Tables 1 and 2 compare the four different approaches at t+6 and t+12, respectively. We also look at how each of the neural network-based approaches performs with and without phases in the input, as well as how each performs using NN or RNN as the algorithm. The results for M3 use class weights since the RNN was only tested with class weights for that approach.

Table 1. Comparison of the four approaches predicting intensity at t+6 (Underlined values are the best within each neural network column, bold values are the best value for each metric, and values in parentheses are standard deviations.)

		MAE		Lag	
	Persistent	0.419		6.000	
	Posner	1.531		**4.000**	
Input	Approach	NN	RNN	NN	RNN
No phases	M1	0.441 (0.018)	**<u>0.379</u>** (0.025)	<u>4.444</u> (0.396)	5.222 (0.396)
	M3	<u>0.432</u> (0.015)	0.433 (0.042)	5.211 (0.654)	6.133 (0.665)
Phases	M1	0.475 (0.012)	0.405 (0.033)	5.289 (0.512)	<u>4.589</u> (0.655)
	M3	0.448 (0.023)	0.470 (0.026)	4.500 (0.846)	4.744 (0.547)

Table 2. Comparison of the four approaches predicting intensity at t+12 (Underlined values are the best within each neural network column, bold values are the best value for each metric, and values in parentheses are standard deviations.)

		MAE		Lag	
	Persistent	0.732		12.000	
	Posner	1.586		**6.722**	
Input	Approach	NN	RNN	NN	RNN
No phases	M1	0.690 (0.037)	**<u>0.599</u>** (0.016)	9.600 (0.422)	9.722 (0.798)
	M3	<u>0.652</u> (0.020)	0.680 (0.036)	10.111 (0.674)	10.367 (0.302)
Phases	M1	0.653 (0.036)	0.648 (0.041)	<u>7.956</u> (0.523)	8.733 (0.577)
	M3	0.677 (0.024)	0.700 (0.018)	8.922 (0.567)	<u>8.489</u> (1.191)

In Table 1, one of the bold values and three out of four of the underlined values are from M1 as the approach, which is unexpected since the M3 approach was designed to minimize errors. Posner's forecasting matrix method has a bold value for lag, which will be explained later in this section. Table 2 also has a bold value in an M1 row, but the underlines are divided evenly between M1 and M3, rather than M1 having most of the underlines as in Table 1. Again, Posner's method has the bold value for lag. All of the best results outperform those of the persistent model.

Looking at the MAE columns, M3 with NN and M1 with RNN give the best results for for both t+6 and t+12, both without phases. This shows that the MAE of each approach depends on which neural network is used, but either way, including phases does not yield significant improvement. The MAE values obtained using Posner's forecasting matrix method are much higher compared to the MAE values using neural networks. However, for t+6, only two neural network models perform better than the persistent model in terms of MAE, with only one of those two models being significantly better. For t+12, almost all neural network models have significantly lower MAE than the persistent model. This can be explained by the fact that the neural network models have more fluctuations in their predictions, which can yield larger vertical gaps, particularly when the proton flux is relatively flat or slow-rising.

For the lag, starting with the neural network-based approaches, M1 has the best value in three out of four cases, with M3 having the lowest lag with RNN and phase inputs when predicting t+12. Phase inputs help to improve the lag in three out of four cases as well. Since the persistent model's prediction is equal to the current value, the lag is exactly 30 minutes when predicting 30 minutes ahead, and the lag is exactly 1 hour when predicting 1 hour ahead. The persistent model's lags do not outperform any of the other methods. Posner's forecasting matrix method yields the lowest lag; however, this is mainly because the method generally overpredicts the proton intensity, as will be illustrated in Section 4.3.1.

For further visualization and analysis, we must select a generally better method out of the neural network-based methods. For t+6, M1 with RNN and no phases has the lowest MAE, but this method has a relatively high lag. The next best-performing method

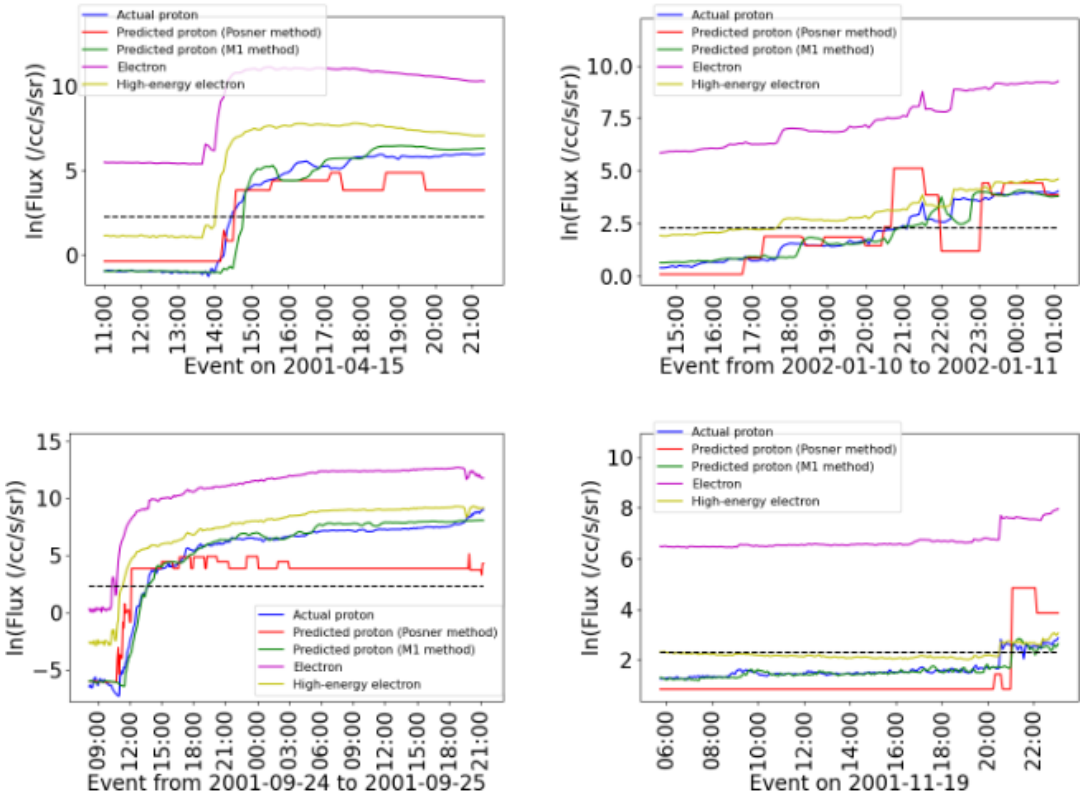


Figure 4. Plots of event predictions using Posner's forecasting matrix method (red) and M1 method with RNN and phase inputs (green), predicting t+6. The arrows point out the times of crossing the ln(10) threshold using M1.

in terms of MAE is M1 with RNN and phases; the metrics for this method are not significantly different from the best-performing metrics, so it can be considered the method with the best balance of both low MAE and low lag. A similar assessment can be made for t+12, except that the lag for M1 with RNN and phases is not significantly better than Posner's method. Therefore, the remainder of the paper will focus on M1 with RNN and phases.

Additionally, the best neural network-based method will be compared with one of the two baseline approaches. Analyzing the lag results, it can be seen that the persistent model is a poor choice for this problem because it is unable to predict SEP events ahead of time, which is the goal of this paper. For a t+6 forecasting example, if an SEP occurs at time t, the persistent model will predict that an SEP will start at time t+6. With the advance warning described in Section 4.1.2, we can start an alert at time t. However, the alert is not useful because the SEP event has already started at time t. Alert 5 in Figure 3 illustrates this issue, where the advance warning (yellow bar) starts at the same time as the SEP event (blue bar). Therefore, the persistent model cannot forecast correctly, and will not be included in further discussion. Posner's method has high MAE values, but its low lags allow it to predict SEP events early, so it will be used as the baseline for comparison against the best neural network-based method.

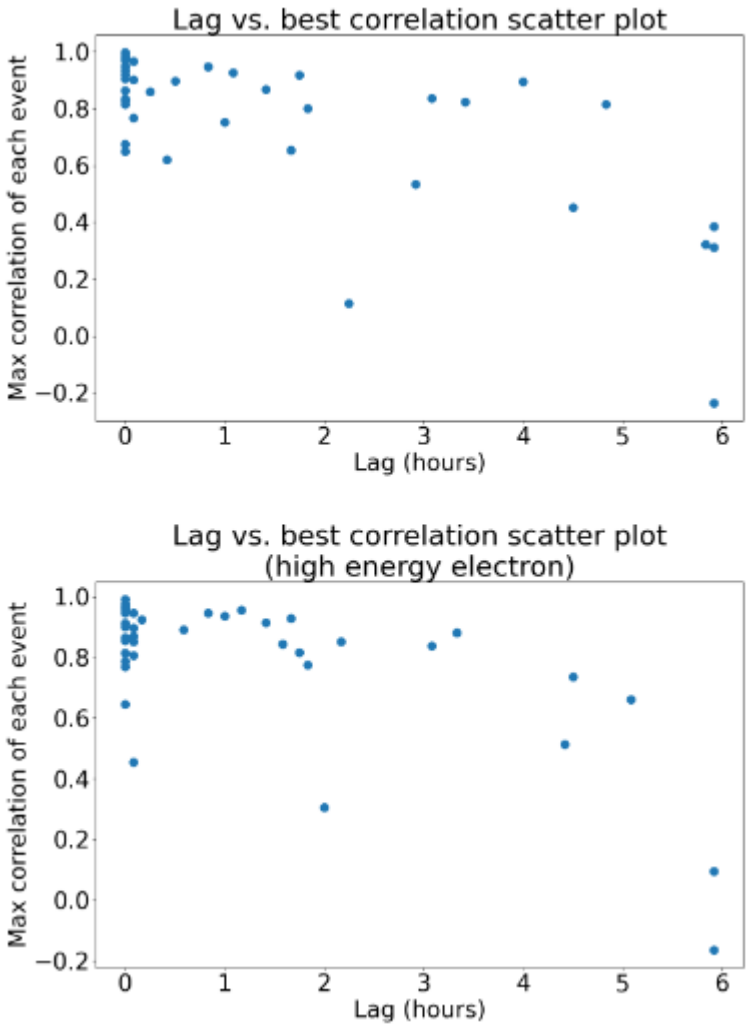


Figure 5. Comparison of lag and max correlation for each event between electron and proton (top) and between high energy electron and proton (bottom).

4.3.1 Sample Prediction Plots

In addition to the tables, prediction plots for four representative events are shown in Figure 4, using the M1 approach with the overall best results as seen in Table 1, as well as Posner's method. The upper-left plot is a relatively-fast rising event, while the upper-right is a slower-rising event. The lower-left reaches very high intensities, while the lower-right is just barely above the threshold of ln(10). Each of the plots shows three hours of background before the event, which are not used during evaluation. With the M1 method, these plots show fairly small errors on both the x- and y-axes, which is consistent with the results in Table 1 for MAE and lag. Posner's forecasting matrix method generally tends to overpredict the proton intensity when the electron intensity has a sharp rise and underpredict when the electron intensity is flat. Overpredicting the rising edge can cause the predictions to be ahead of the actual rising edge, resulting in a smaller lag on average.

4.3.2 Analysis of Results

The plots in Figure 5 can help to better understand why it is difficult to achieve low lag in the predictions for some events. To generate these plots, a similar procedure is done for calculating lag between the actual and predicted values, but this is done for electron and proton values instead, and uses Pearson correlation rather than MAE to choose the best lag since the actual electron and proton values cannot be compared with each other. Additionally, correlations are calculated for all 5-minute shifts up to 6 hours, rather than only 2 hours. Out of the 39 events in the data set, there are 15 events (38%) with zero lag between the proton intensity and >0.25 MeV electron intensity, 5 events (13%) with between 5 minutes and 1 hour of lag, and 19 events (49%) with more than one hour of lag. For the >0.67 MeV electron channel, there are 15 events (38%) with zero lag, 10 events (26%) with up to 1 hour of lag, and 14 events (36%) with more than one hour of lag. Because 51% and 64% of events (using the >0.25 MeV and >0.67 MeV electron channels, respectively) have a lag of an hour or less, there is limited information that can be used ahead of the SEP event for prediction, and our forecasts are likely to have a lag. The large number of SEP events with less than an hour of lag is visualized by the clustering of dots towards the left of the plots in Figure 5. It would be more desirable if more of the dots were clustered toward the upper-right of the plots, as this would allow for earlier predictions than what the algorithm is currently capable of, while the electron and proton are still highly correlated. However, despite limitations in the numerical lag results for predicting SEP intensity-time profile, the predictions are still sufficient to warn of the occurrence of an incoming SEP event, as will be demonstrated in the next section.

4.4 Results of SEP Event Forecasting

We compare our method (M1 with RNN and phase inputs) with Posner's method according to the 4 approaches (W, EW, AW, and EAW) to extended and advanced warnings described in Section 4.1.2. The results in terms of event forecasting metrics are shown in Tables 3 and 4. These results use 2 hours of extended warning, which is found by experimenting with different durations. By varying the extended warning duration, Figure 6 shows the trend of F1 scores for each method. In most cases, the F1 scores plateau around 2 hours.

In approach W, which uses neither the advance nor extended warnings, Posner's method yields more true positives than M1, but also more false positives. M1 yields few true positives. Posner's method has both better recall and precision despite the high number of false positives, and therefore a higher F1 score.

In approach EW, which includes the extended warning but no advance warning, M1 has more true positives than in approach W, while Posner's method remains the same in terms of true positives. For both methods, the number of false positives drops compared to approach W, with M1 still having fewer false positives than Posner; this results in a higher precision for M1. When predicting at t+6, Posner's method has more true positives than when predicting t+12, so this difference results in Posner having a higher F1 score than M1 at t+6, but lower than M1 at t+12.

In approach AW, which includes the advance warning but no extended warning, the number of true positives predicted by M1 increases substantially compared to not using advance warning. At t+6, M1 has slightly fewer true positives than Posner, but slightly more at t+12. The number of false positives remains lower for M1 than for Posner. Since the numbers of true positives are close for both methods, the recall values are close as well. M1 having fewer false positives results in a slightly higher precision for M1 than for Posner, therefore resulting in M1 having a slightly better F1 score.

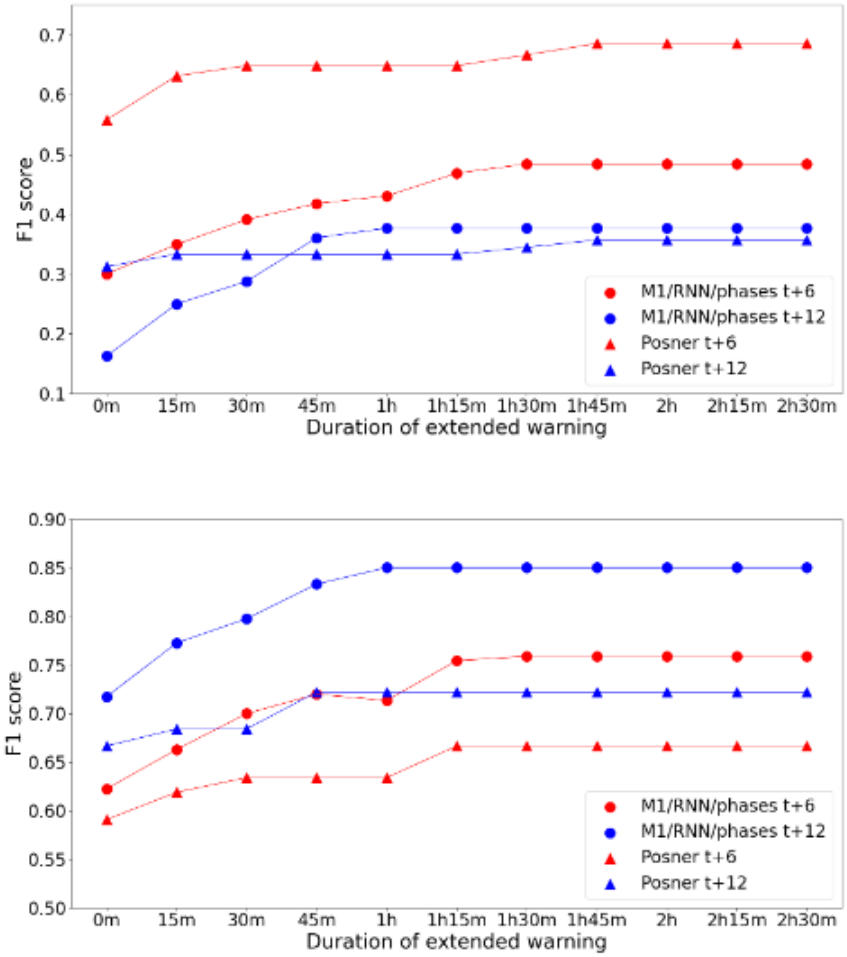


Figure 6. Comparison between our method and Posner's method in event forecasting with different durations of extended warning. The top plot does not include advance warning, using Approach W when the extended warning duration is zero and Approach EW when the extended warning duration is above zero. The bottom plot includes advance warning, using Approach AW when the extended warning duration is zero and Approach EAW when the extended warning duration is above zero.

Finally, approach EAW, in which both the advance and extended warning are included, shows similar trends to those of approach AW. The number of true positives predicted by M1 increase even more from approach AW, and the numbers of false positives decrease from approach AW, with a larger decrease for M1 than for Posner. As in approach AW, M1 has fewer true positives than Posner at t+6, but more at t+12, resulting in a slightly lower recall for M1 at t+6, but higher at t+12. Due to the very low numbers of false positives for M1 compared to Posner, M1 has much higher precision values. These result in M1 having higher F1 scores compared to Posner.

From these results, we generally observe that M1 performs better than Posner's method when advance warning is included. One reason is that M1 generally lags the rising edge as observed in Section 4.3.1, and advance warning helps mitigate the lag in M1 and predicts more positives. Also, we observe that Posner's method yields more true positives even without advance warning. One reason is that Posner's method tends to over-

Table 3. Results of the four event forecasting applications at t+6. (Underlined F1 values are the best within each approach, and the bold F1 value is the best across all approaches.)

	Approach W		Approach EW		Approach AW		Approach EAW	
Method	M1	Posner	M1	Posner	M1	Posner	M1	Posner
TP	4.8	12.0	6.2	12.0	12.2	13.0	12.6	13.0
FN	13.2	6.0	11.8	6.0	5.8	5.0	5.4	5.0
FP	9.2	13.0	1.4	5.0	9.0	13.0	2.6	8.0
Recall	0.27	0.67	0.34	0.67	0.68	0.72	0.70	0.72
Precision	0.34	0.48	0.82	0.71	0.58	0.50	0.83	0.62
F1	0.30	0.56	0.48	0.69	0.62	0.59	**0.76**	0.67

Table 4. Results of the four event forecasting applications at t+12. (Underlined F1 values are the best within each approach, and the bold F1 value is the best across all approaches.)

	Approach W		Approach EW		Approach AW		Approach EAW	
Method	M1	Posner	M1	Posner	M1	Posner	M1	Posner
TP	2.2	5.0	4.6	5.0	13.2	13.0	14.2	13.0
FN	15.8	13.0	13.4	13.0	4.8	5.0	3.8	5.0
FP	6.8	9.0	1.8	5.0	5.6	8.0	1.2	5.0
Recall	0.12	0.28	0.26	0.28	0.73	0.72	0.79	0.72
Precision	0.24	0.36	0.72	0.50	0.70	0.62	0.92	0.72
F1	0.16	0.31	0.38	0.36	0.72	0.67	**0.85**	0.72

predict the proton intensities, and hence tends to be ahead of the rising edge. However, this also results in Posner's method having more false positives, which occur in all four approaches. Furthermore, for both methods, we observe that having neither the advance warning nor extended warning yields the lowest F1 scores, while having both advance and extended warning yields the highest F1 scores, as indicated in Tables 3 and 4.

5 Conclusions

The problem studied in this work is the forecasting of future proton flux given a time series of past and current electron and proton flux. We use a single model as the basic approach, and present another approach which splits the data by intensity ranges and select the model using a separate machine learning model. We compare regular neural networks and recurrent neural networks, and experiment with phase inputs.

Overall, our results indicate that a single RNN model generally performs better in terms of proton flux prediction, less MSE in predicting proton flux, but a larger lag, than the forecasting matrix method proposed by Posner (2007). Based on the direct prediction of SEP proton intensity and timing, the RNN model makes fewer true positive and false positive predictions of SEP proton events than Posner, yielding poor recall, precision, and F1 scores. This is because the RNN prediction of SEP proton intensity is gen-

erally delayed. Suppose we allow warnings to be issued immediately after predictions are made and extended by up to 2 hours of predicted duration. In that case, the RNN model prediction of SEP proton events can dramatically improve the true positive while the false positive remains low, which results in better recall, precision, and F1 scores.

Based on our analysis of the electron and proton time series, obtaining a lag of zero in our results is quite difficult with just electron and proton time series, and other features preceding the proton event would be required in order to achieve a lag of zero between the predicted and actual future proton values. The prediction could also be improved by using a longer dataset; our current data does not cover an entire solar cycle, and ends on a solar maximum. This means that the algorithm is trained on few events and evaluated on many events; the algorithm would be more effective when the distributions of events in the training and test sets are similar.

Open Research

This work uses data obtained from Müller-Mellin et al. (1995) [1]. The SOHO/EPHIN project is supported under grant No. 50 OC 0105 by the German Bundesminister für Wirtschaft through the Deutsches Zentrum für Luft- und Raumfahrt (DLR).

Acknowledgments

This work was supported in part by NASA under Grants 80NSSC20K0298, 80NSSC19K0076, 80NSSC18K0644, 80NSSC20K0286, and 80NSSC21K0004. We want to thank Arik Posner for helpful discussion on his forecasting matrix method used for comparison with this work.

References

Balch, C. C. (2008). Updated verification of the space weather prediction center's solar energetic particle prediction model. *Sp. Weather*, *6*(1), 1–13. doi: 10.1029/2007SW000337

Belov, A. (2009). Properties of solar X-ray flares and proton event forecasting. *Adv. Sp. Res.*, *43*(4), 467–473. Retrieved from `http://dx.doi.org/10.1016/j.asr.2008.08.011` doi: 10.1016/j.asr.2008.08.011

Bobra, M. G., & Ilonidis, S. (2016). Predicting Coronal Mass Ejections Using Machine Learning Methods. *Astrophys. J.*, *821*(2), 1–7. Retrieved from `http://arxiv.org/abs/1603.03775{\%}0Ahttp://dx.doi.org/10.3847/0004-637X/821/2/127` doi: 10.3847/0004-637X/821/2/127

Cane, H. V., Richardson, I. G., & von Rosenvinge, T. T. (2010). A study of solar energetic particle events of 1997–2006: Their composition and associations. *J. Geophys. Res.*, *115*(August 2009), 1–18. doi: 10.1029/2009JA014848

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Dierckxsens, M., Tziotziou, K., Dalla, S., Patsou, I., Marsh, M. S., Crosby, N. B., ... Tsiropoula, G. (2015). *Relationship between Solar Energetic Particles and Properties of Flares and CMEs: Statistical Analysis of Solar Cycle 23 Events* (Vol. 290) (No. 3). doi: 10.1007/s11207-014-0641-4

Garcia, H. (2004). Forecasting methods for occurrence and magnitude of proton storms with solar hard X rays. *Space Weather*, *2*(6).

Georgoulis, M. K. (2008). Magnetic complexity in eruptive solar active regions and

[1] `http://www2.physik.uni-kiel.de/SOHO/phpeph/EPHIN.htm`

associated eruption parameters. *Geophys. Res. Lett.*, *35*(6), 5–9. doi: 10.1029/2007GL032040

Huang, X., Wang, H., Xu, L., Liu, J., Li, R., & Dai, X. (2018). Deep Learning Based Solar Flare Forecasting Model. I. Results for Line-of-sight Magnetograms. *Astrophys. J.*, *856*(1), 7. Retrieved from `http://dx.doi.org/10.3847/1538-4357/aaae00` doi: 10.3847/1538-4357/aaae00

Huang, X., Wang, H. N., & Li, L. P. (2012). Ensemble prediction model of solar proton events associated with solar flares and coronal mass ejections. *Res. Astron. Astrophys.*, *12*(3), 313–321. doi: 10.1088/1674-4527/12/3/007

Kahler, S., Cliver, E., & Ling, A. (2005). Validating the proton prediction system. *AGUSM*, *2005*, SH41A–03.

Laurenza, M., Cliver, E. W., Hewitt, J., Storini, M., Ling, A. G., Balch, C. C., & Kaiser, M. L. (2009). A technique for short-term warning of solar energetic particle events based on flare location, flare size, and evidence of particle escape. *Sp. Weather*, *7*(4), 1–18. doi: 10.1029/2007SW000379

Müller-Mellin, R., Kunow, H., Fleißner, V., Pehlke, E., Rode, E., Röschmann, N., ... Henrion, J. (1995, December). COSTEP - Comprehensive Suprathermal and Energetic Particle Analyser. , *162*(1-2), 483-504. doi: 10.1007/BF00733437

Núñez, M. (2011). Predicting solar energetic proton events (E>10 MeV). *Space Weather*, *9*(7).

Park, E., Moon, Y.-J., Shin, S., Yi, K., Lim, D., Lee, H., & Shin, G. (2018). Application of the Deep Convolutional Neural Network to the Forecast of Solar Flare Occurrence Using Full-disk Solar Magnetograms. *Astrophys. J.*, *869*(2), 91. Retrieved from `http://dx.doi.org/10.3847/1538-4357/aaed40` doi: 10.3847/1538-4357/aaed40

Park, J., Moon, Y. J., & Gopalswamy, N. (2012). Dependence of solar proton events on their associated activities: Coronal mass ejection parameters. *J. Geophys. Res. Sp. Phys.*, *117*(8), 1–7. doi: 10.1029/2011JA017477

Posner, A. (2007). Up to 1-hour forecasting of radiation hazards from solar energetic ion events with relativistic electrons. *Sp. Weather*, *5*(5), 1–28. doi: 10.1029/2006SW000268

Reames, D. V. (2013). The two sources of solar energetic particles. *Space Sci. Rev.*, *175*(January), 53–92. doi: 10.1007/s11214-013-9958-9

Richardson, I., Mays, M., & Thompson, B. (2018). Prediction of solar energetic particle event peak proton intensity using a simple algorithm based on CME speed and direction and observations of associated solar phenomena. *Space Weather*, *16*(11), 1862–1881.

Shen, C., Wang, Y., Ye, P., Zhao, X. P., Gui, B., & Wang, S. (2007). Strength of Coronal Mass Ejection-driven Shocks near the Sun and Their Importance in Predicting Solar Energetic Particle Events. *Astrophys. J.*, *670*(1), 849–856. doi: 10.1086/521716

Smart, D. F., & Shea, M. A. (1976). PPS76 - A Computerized "Event Mode" Solar Proton Forecasting Technique. *Sol. Terrest Predict. Proc.*, 406.

Smart, D. F., & Shea, M. A. (1989). PPS-87: A new event oriented solar proton prediction model. *Adv. Sp. Res.*, *9*(10), 281–284. doi: 10.1016/0273-1177(89)90450-X

St. Cyr, O. C., Posner, A., & Burkepile, J. T. (2017). Solar energetic particle warnings from a coronagraph. *Sp. Weather*, *15*(1), 240–257. doi: 10.1002/2016SW001545

Torres, J. (2020). *A machine learning approach to forecasting sep events with solar activities* (Master's thesis, Florida Institute of Technology). `https://repository.lib.fit.edu/handle/11141/3215`.

Figure 1.

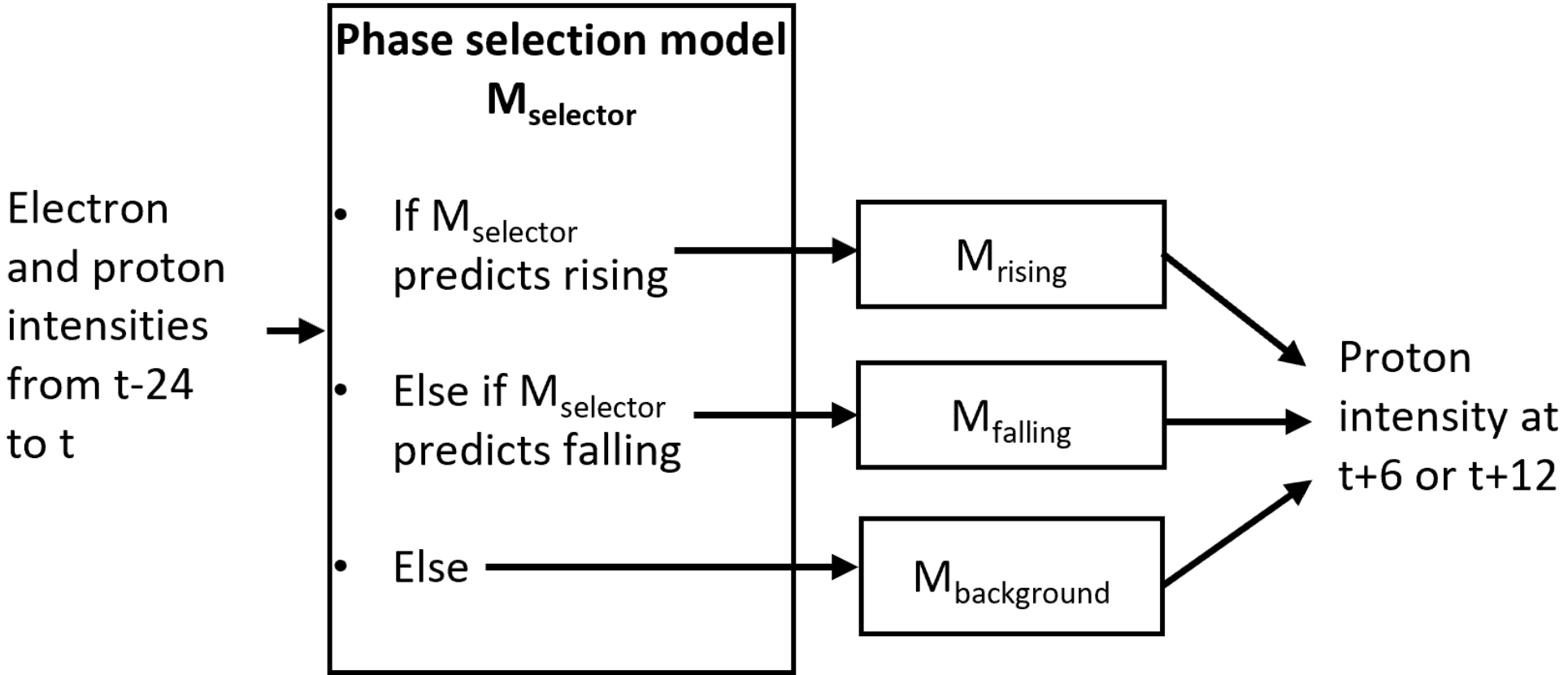

Phase selection model
$M_{selector}$
Electron and proton intensities from t-24 to t
If $M_{selector}$ predicts rising
Else if $M_{selector}$ predicts falling
Else
M_{rising}
$M_{falling}$
$M_{background}$
Proton intensity at t+6 or t+12

Figure 2.

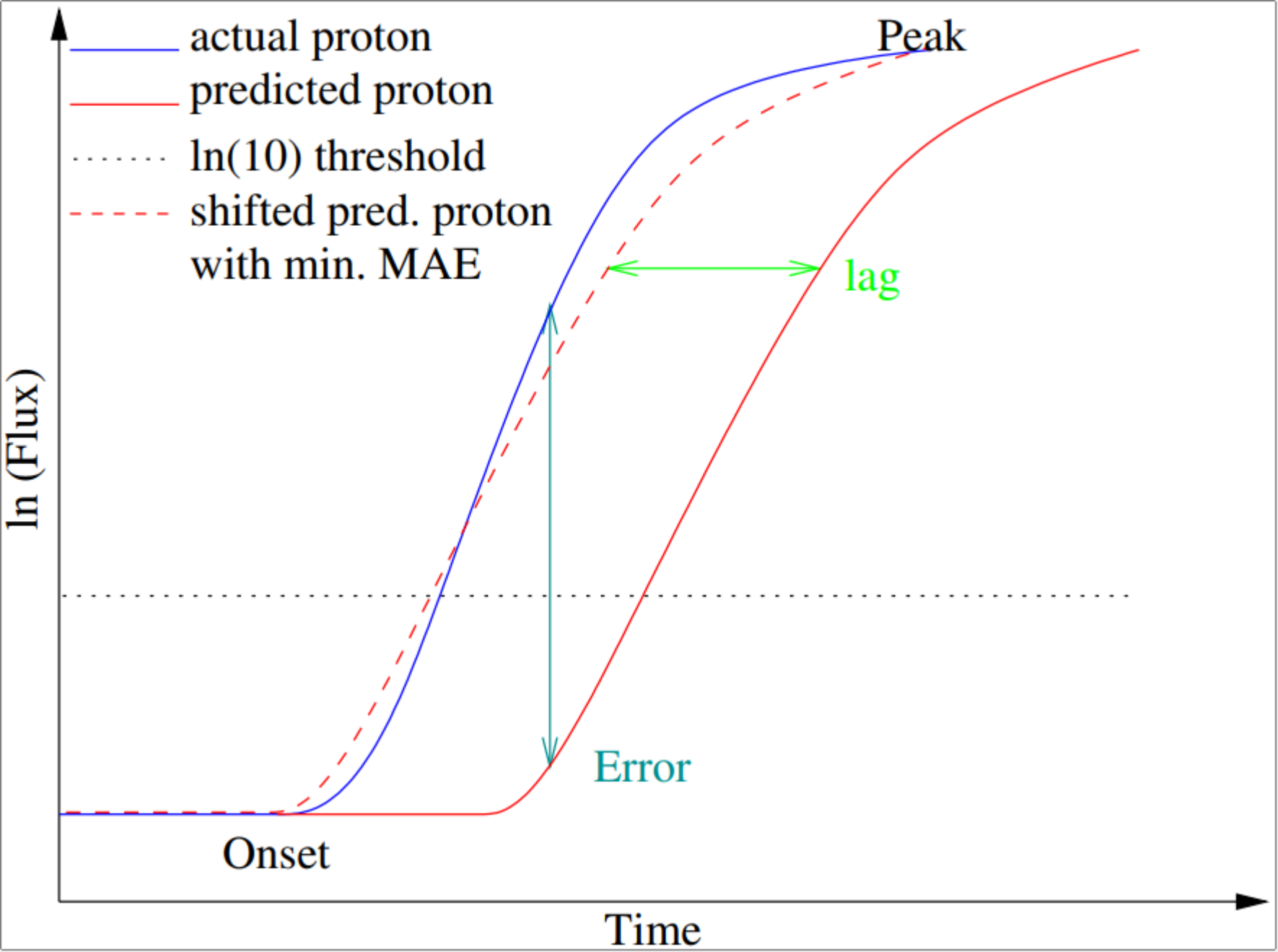

actual proton
predicted proton
ln(10) threshold
shifted pred. proton
with min. MAE
Peak
lag
Error
Onset
ln (Flux)
Time

Figure 3.

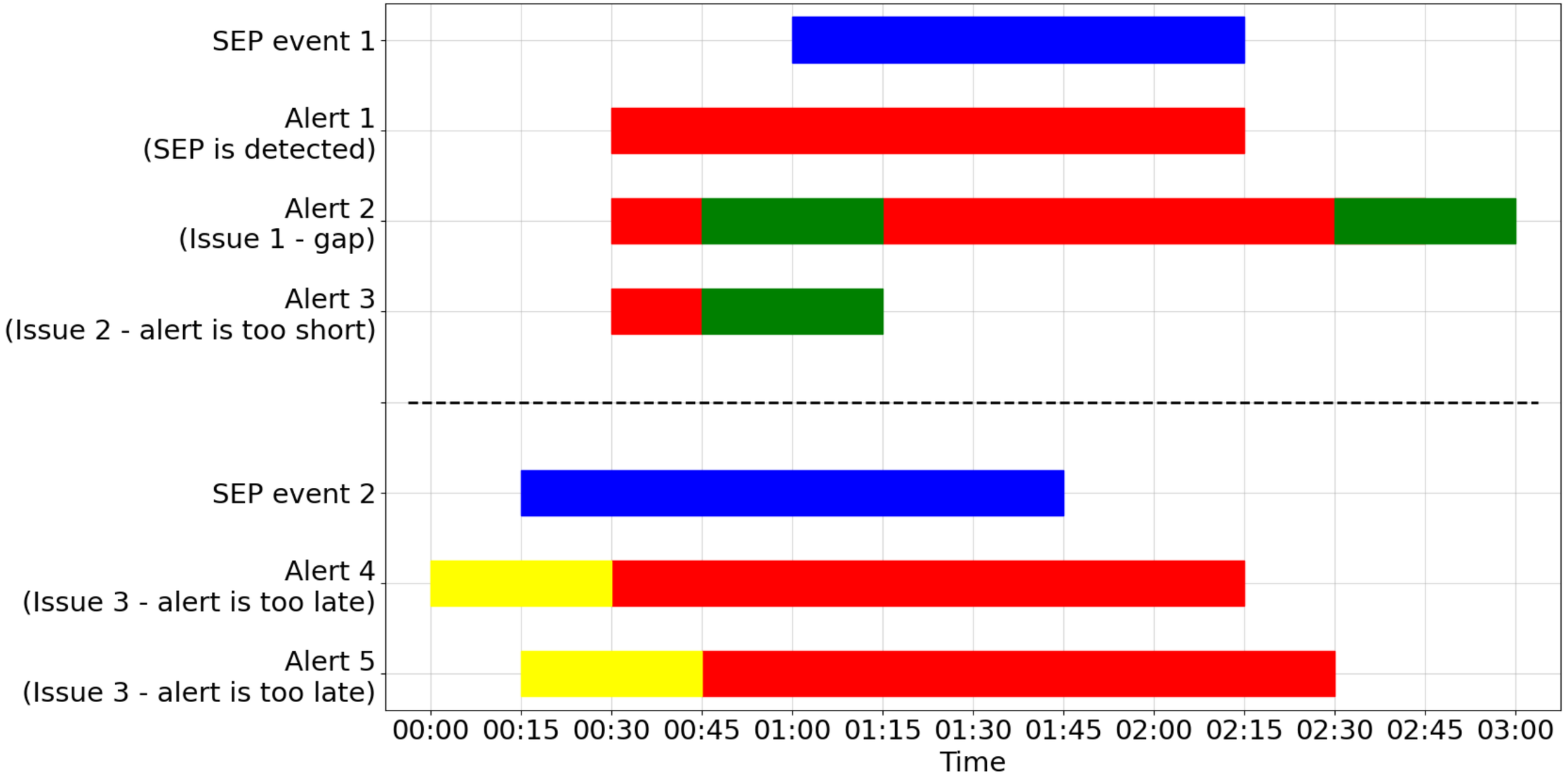
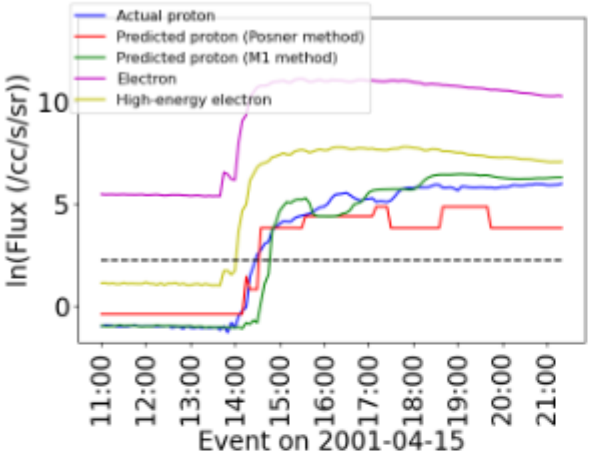
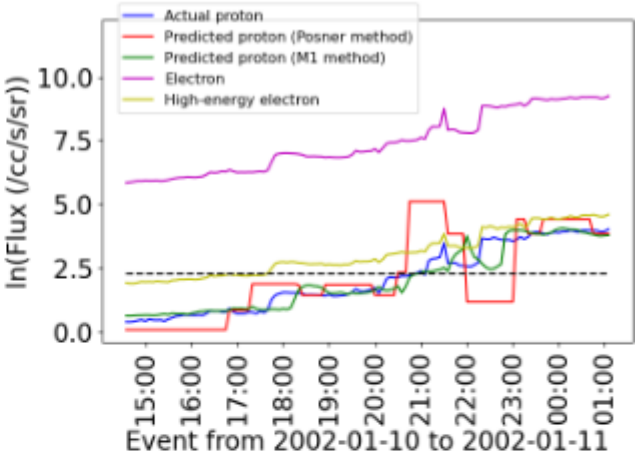
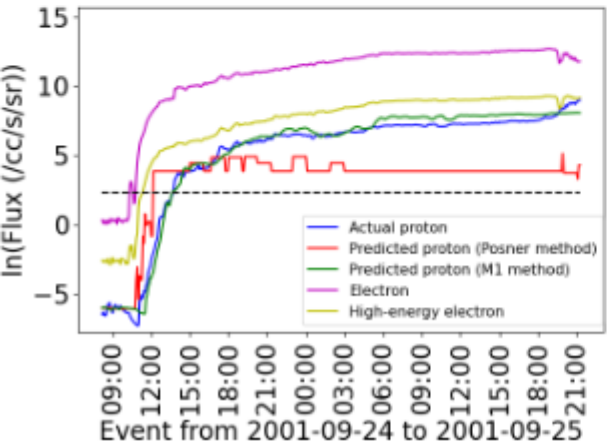

SEP event 1
Alert 1
(SEP is detected)
Alert 2
(Issue 1 - gap)
Alert 3
(Issue 2 - alert is too short)
SEP event 2
Alert 4
(Issue 3 - alert is too late)
Alert 5
(Issue 3 - alert is too late)
00:00
00:15
00:30
00:45
01:00
01:15
01:30
01:45
02:00
02:15
02:30
02:45
03:00
Time

Figure 4.


Actual proton
Predicted proton (Posner method)
Predicted proton (M1 method)
Electron
High-energy electron
ln(Flux (/cc/s/sr))
10
5
0
11:00
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
Event on 2001-04-15


Actual proton
Predicted proton (Posner method)
Predicted proton (M1 method)
Electron
High-energy electron
ln(Flux (/cc/s/sr))
10.0
7.5
5.0
2.5
0.0
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
00:00
01:00
Event from 2002-01-10 to 2002-01-11


ln(Flux (/cc/s/sr))
15
10
5
0
−5
Actual proton
Predicted proton (Posner method)
Predicted proton (M1 method)
Electron
High-energy electron
09:00
12:00
15:00
18:00
21:00
00:00
03:00
06:00
09:00
12:00
15:00
18:00
21:00
Event from 2001-09-24 to 2001-09-25

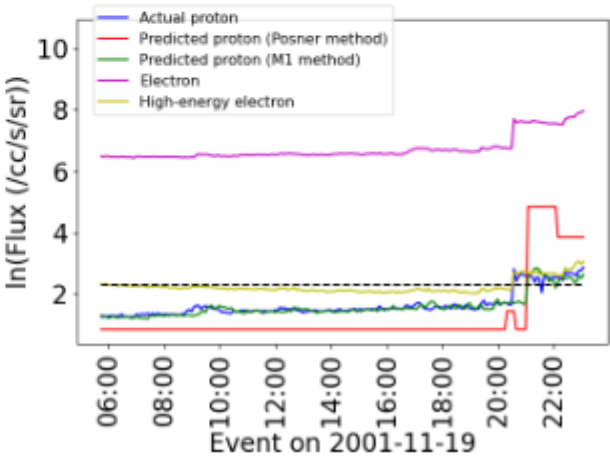
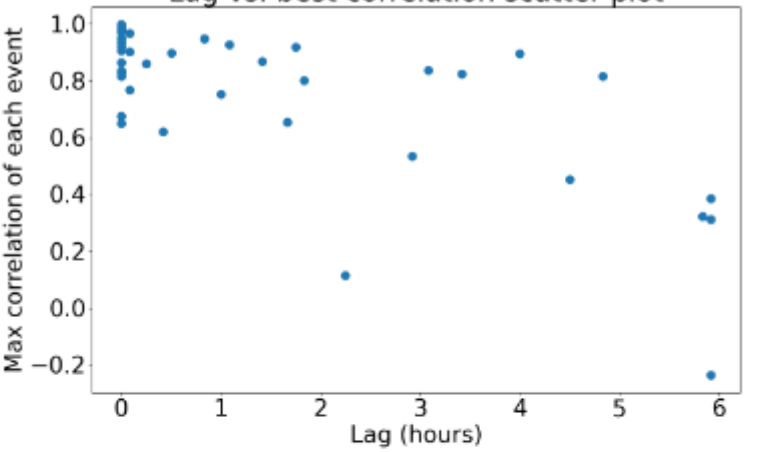

Actual proton
Predicted proton (Posner method)
Predicted proton (M1 method)
Electron
High-energy electron
ln(Flux (/cc/s/sr))
10
8
6
4
2
06:00
08:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Event on 2001-11-19

Figure 5.


Lag vs. best correlation scatter plot
Max correlation of each event
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
1
2
3
4
5
6
Lag (hours)

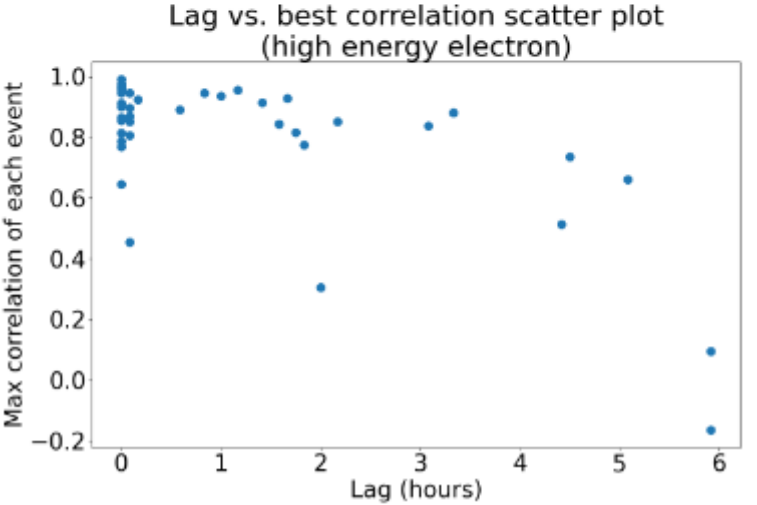
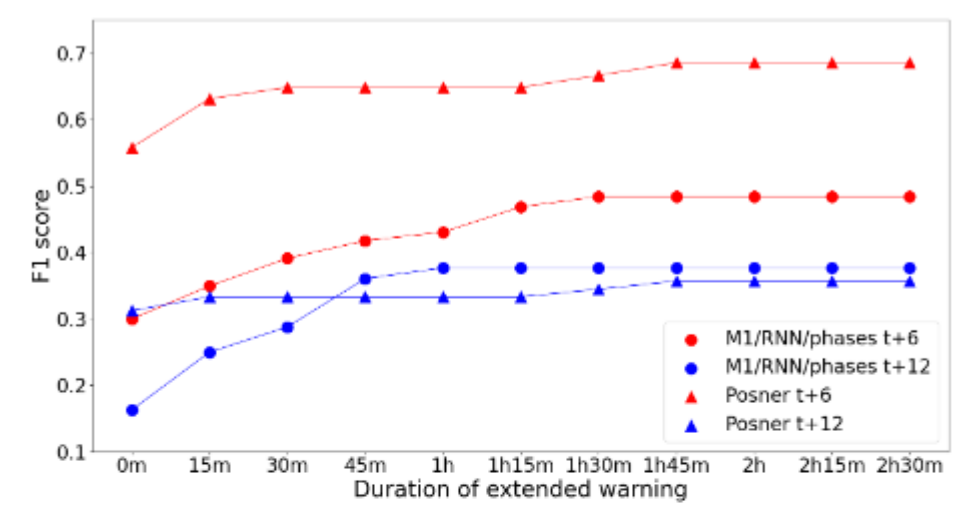
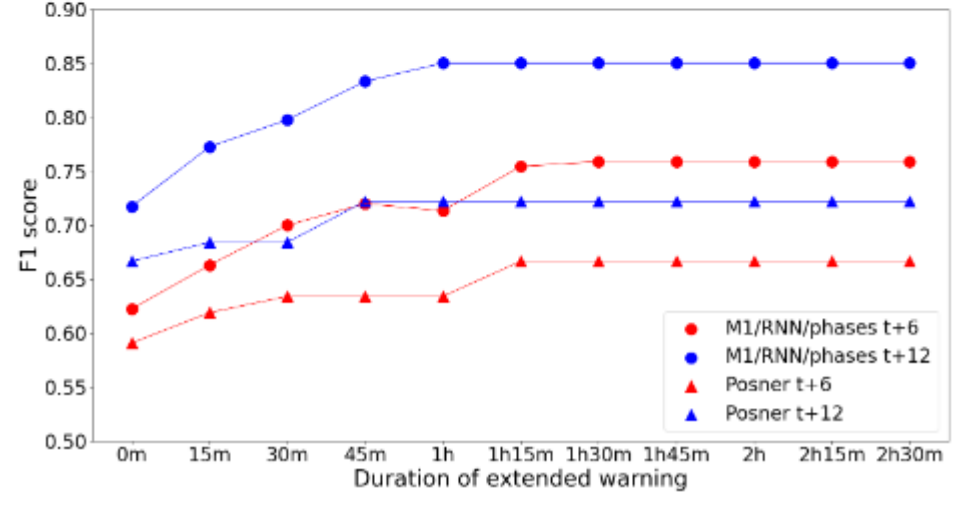

Lag vs. best correlation scatter plot
(high energy electron)
Max correlation of each event
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
1
2
3
4
5
6
Lag (hours)

Figure 6.


F1 score
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0m
15m
30m
45m
1h
1h15m
1h30m
1h45m
2h
2h15m
2h30m
Duration of extended warning
M1/RNN/phases t+6
M1/RNN/phases t+12
Posner t+6
Posner t+12


F1 score
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0m
15m
30m
45m
1h
1h15m
1h30m
1h45m
2h
2h15m
2h30m
Duration of extended warning
M1/RNN/phases t+6
M1/RNN/phases t+12
Posner t+6
Posner t+12